Statistics: A refresher

Mariano Mendez
Kapteyn Astronomical Institute
University of Groningen
The Netherlands

Probability: Definitions

X, Y, Z = (sets) of events, either discrete or continuous

$\sim X, \sim Y, \sim Z$ = complement of the same sets of events (**negation**)

$P(X)$ = Probability of X (**Probability Distribution Function**, **PDF**, of X)

$P(X|Y)$ = Probability of X given Y (**Conditional probability**)

$Z = X \; or \; Y$ = set of events that belong to X, Y, or both (**Union**)

$Z = X \; and \; Y$ = set of events that belong both to X and Y (**Intersection**)

with $P(false) = 0$ and $P(true) = 1$, defining certainty.

Probability: Rules

$P(\sim X) = 1 - P(X)$

$P(X \text{ and } Y) = P(X,Y) = P(X \mid Y)\, P(Y)$

If X and Y are independent $\Rightarrow$ $P(X \mid Y) = P(X)$

$P(X,Y) = P(X)\, P(Y)$

$P(X \text{ or } Y) = P(X) + P(Y) - P(X,Y)$

If X and Y are mutually exclusive $\Rightarrow$ $P(X,Y) = 0$.

Probability: Bayes theorem

$$P(X,Y) = P(Y,X) \Rightarrow P(X \mid Y)\, P(Y) = P(Y \mid X)\, P(X)$$

While this is an innocent-looking formula, it is the source of heated debates in science.

The importance of this theorem comes from the interpretation given to this formula.

E.g., if we call X **= model**, Y **= data.**

Probability: Bayes theorem

$P(X \mid Y)\, P(Y) = P(Y \mid X)\, P(X)$

X = **model**; Y = **data**

$P(data|model)$: probability of the data given the model = **Likelihood**

$P(model|\textbf{data})$: probability of the model given the data = **Posterior**

$P(model)$: probability of the model before the experiment = **Prior**

$P(data)$: a normalization such that $\int P(X \mid Y)\, dX = 1$ = **Evidence**

Probability

This is a conceptual revolution. For **"Bayesians"**, a probability represents a ***degree-of-belief*** or ***plausibility***: how much one thinks that something (e.g., a model) is true, based on the evidence (i.e., data) at hand.

To the 19th century mathematicians this seemed too vague and subjective an idea to be the basis of a rigorous mathematical theory. So they redefined probability as the long-run relative ***frequency*** with which an event occurred, given ***infinitely*** many (hypothetically) repeated (experimental) trials. Since frequencies can be measured, probability could then be seen as an objective tool for dealing with random phenomena. This is the so-called **"Frequentist"** approach.

Probability

Bayesian theorem is more profound than just a mathematical formula, since it provides a way to describe the way we reason.

1. We have a belief about something (=Prior).
2. We carry out an experiment to test our belief (=Likelihood).
3. We adjust our belief based on the result of the experiment (=Posterior).
   Bayes theorem tells us how to do this.
4. Our new belief becomes the new Prior, and we go back to 1.

$$P(\textbf{model} \mid \textbf{data}) \propto P(\textbf{data} \mid \textbf{model}) \times \textbf{Prior}$$

Probability

Probabilities are **always** conditional

Probability that it rains today $\rightarrow$ given that we are in Potch

$\rightarrow$ given that it is (almost) spring

$\rightarrow$ given that it is cloudy

$\rightarrow$ given that …

Probability to get a 6 in die $\rightarrow$ given that the die is fair

$\rightarrow$ given that the thrower is fair

$\rightarrow$ given that the table's surface is …

Probability to have N photons with energies between E_1 and E_2

$\rightarrow$ given that the source is blackbody

$\rightarrow$ given that I use XMM-Newton …

Poisson and Gauss

Two very important Probability Distribution Functions (PDF):

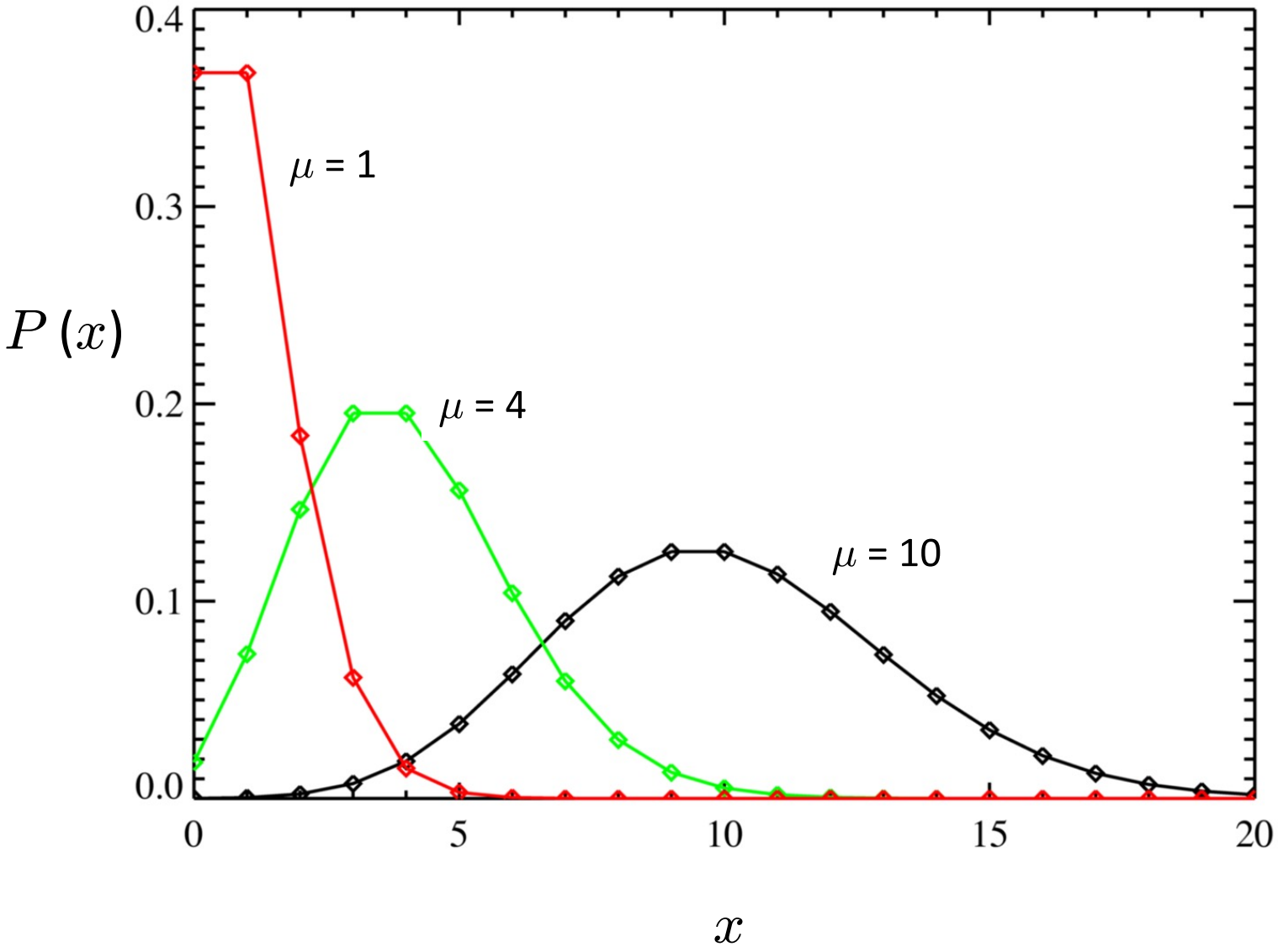
Poissonian distribution (x is a discrete variable):

$$P(x) = \frac{\mu^x}{x!} e^{-\mu}$$

Mean = $$\sum_{x=0}^{\infty} xP(x) = \mu$$

Variance = $$\sum_{x=0}^{\infty} (x-\mu)^2 P(x) = \mu$$

Poisson distribution



Poisson and Gauss

Two very important Probability Distribution Functions (PDF):

Gaussian distribution (x is a continuous variable):

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Mean = $\int xP(x)dx = \mu$

Variance = $\int (x-\mu)^2 P(x)dx = \sigma^2$

Probability

We always measure **random variables**, regardless how accurate our instrument is.

E.g., suppose we count photons from a source that emits a constant photon flux, collected by an instrument within a small time interval Δt and a small energy interval ΔE. (We will ignore the effect of the precision with which we can measure Δt and ΔE.) Since the emission of a photon at the source is independent of whether another photon was already emitted, the emission process is **Poissonian**.

If we repeat the measurement many times, we will not (necessarily) count the same number of photons each time. What is then the "true" photon flux of the source?

Maximum Likelihood

Let us look at the problem of counting photons from the probabilistic point of view.

Suppose that we have a set of N measurements of the number of photons, $\{n_i\}$, i=1,2,…, N, counted within time intervals Δt.

If the distribution of n_i is Poissonian, the probability of measuring n_i photons in interval i **given** that the source emits μ (μ is unknown!!) photons is:

$$P(n_i|\mu) = \frac{\mu^{n_i}}{n_i!}e^{-\mu}$$

Maximum Likelihood

The probability of getting **this set** of N observations $\{n_i\}$, given that the source emits μ photons, if the individual measurements are independent, is (remember the **"and" rule** of probabilities):

$$\mathcal{L} = P(\{n_i\}|\mu) = \prod_{i=1}^{N} P(n_i|\mu) = \prod_{i=1}^{N} \frac{\mu^{n_i}}{n_i!} e^{-\mu}$$

This is called the **Likelihood**. (It is the likelihood of getting the observed dataset given the model.)

The **Principle of Maximum Likelihood (ML)** states that the most likely outcome of an experiment is the one that maximizes $\mathcal{L}$.

It is equivalent (and it is usually easier) to maximize **log** $\mathcal{L}$.

Maximum Likelihood

$$\log \mathcal{L} = \sum_{i=1}^{N} \left[-\mu + n_i \log \mu - \log \left(n_i ! \right) \right]$$

And find μ that maximizes $\log \mathcal{L}$:

$$d \log \mathcal{L} / d\mu = \sum_{i=1}^{N} \left[-1 + n_i / \mu \right] = 0$$

Which yields the well-known result:

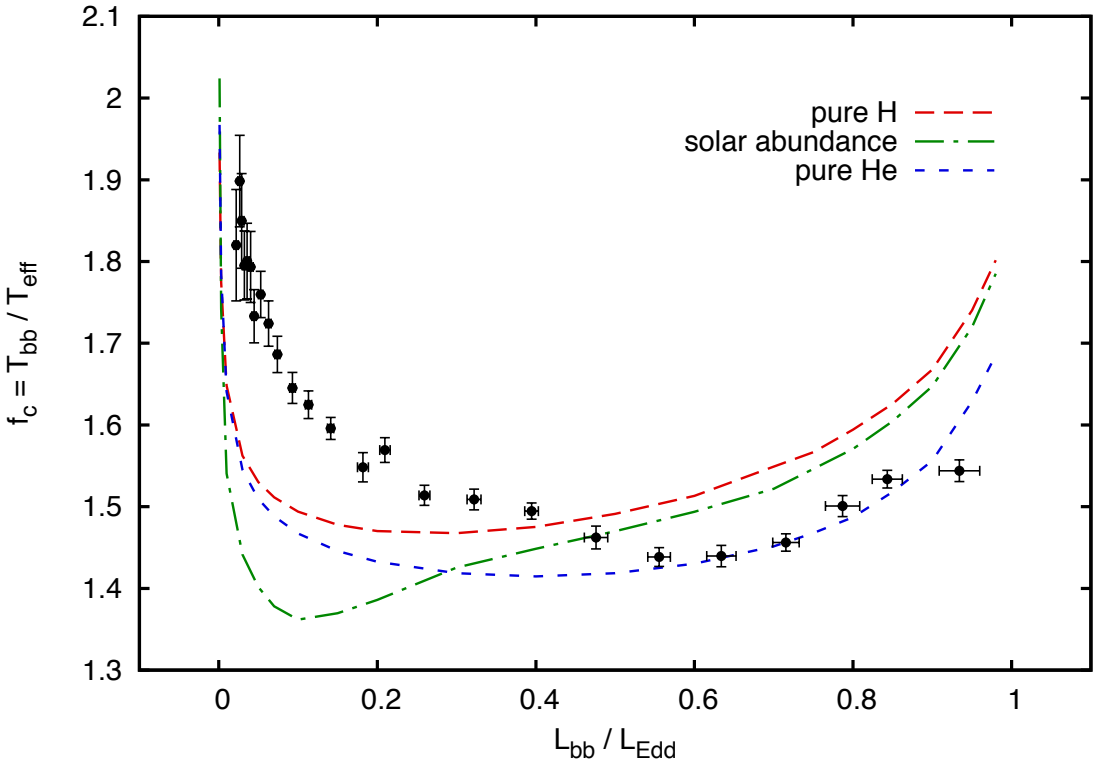
$$\mu = \frac{1}{N} \sum_{i=1}^{N} n_i = \bar{n}_i$$

that the **average** is the ML estimate of the **mean**.

Probability

We always measure **random variables**, regardless how accurate our instrument is. Our measurements will always have an associated error.

Therefore, when we fit a model to these **data**, the **parameters** of the **model** will also be **random variables**.



Parameter estimation

Suppose our data is a set $\{y_i\}$, i = 1,…, N, that represent the spectrum of a source, i.e. the number of photons, y_i, as a function of energy, E_i.

What is the probability of getting **this** spectrum (data) **given an assumed model,** P**(data | model)**?

As in the Poisson example, we need to know the **PDF** of the data given the model. Let us assume that at each energy the data are random variables from a **Gaussian** PDF around the model with errors σ_i.

In other words, at each E_i, the data are a random realization of a model $\boldsymbol{y(E_i\,;\,a)}$, with parameters $\boldsymbol{a}$ ($\boldsymbol{a}$ is a vector of M elements $\boldsymbol{a_1}$, $\boldsymbol{a_2}$ … $\boldsymbol{a_M}$).

The likelihood of the data given the model is:

Parameter estimation

$$\mathcal{L} = P(\{y_i\}|y(E_i;\boldsymbol{a})) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{[y_i - y(E_i;\boldsymbol{a})]^2}{\sigma_i^2}}$$

$$= \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma_i} \right) e^{-\frac{1}{2}\sum_{i=1}^{N}\frac{[y_i - y(E_i;\boldsymbol{a})]^2}{\sigma_i^2}}$$

Maximizing $\mathcal{L}$ with respect to $\boldsymbol{a}$ is equivalent to minimizing

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - y(E_i;\boldsymbol{a})]^2}{\sigma_i^2}$$

If the errors are Gaussian (<u>and only then!!!</u>) Maximum Likelihood is equivalent to minimum χ^2.

Parameter estimation

In reality, the X-ray spectral data are Poissonian (counting photons in energy bins). The χ^2 procedure is therefore not applicable.

However, if **μ is large** the Poisson PDF **tends** to the Gaussian PDF. This is the case when the source is bright and one has many counts per bin.

A common practice is to rebin the data (add together M consecutive energy bins) if the source is weak, to approach the Gaussian regime. This has the disadvantage of losing spectral resolution (narrow emission/absorption lines are diluted in the continuum).

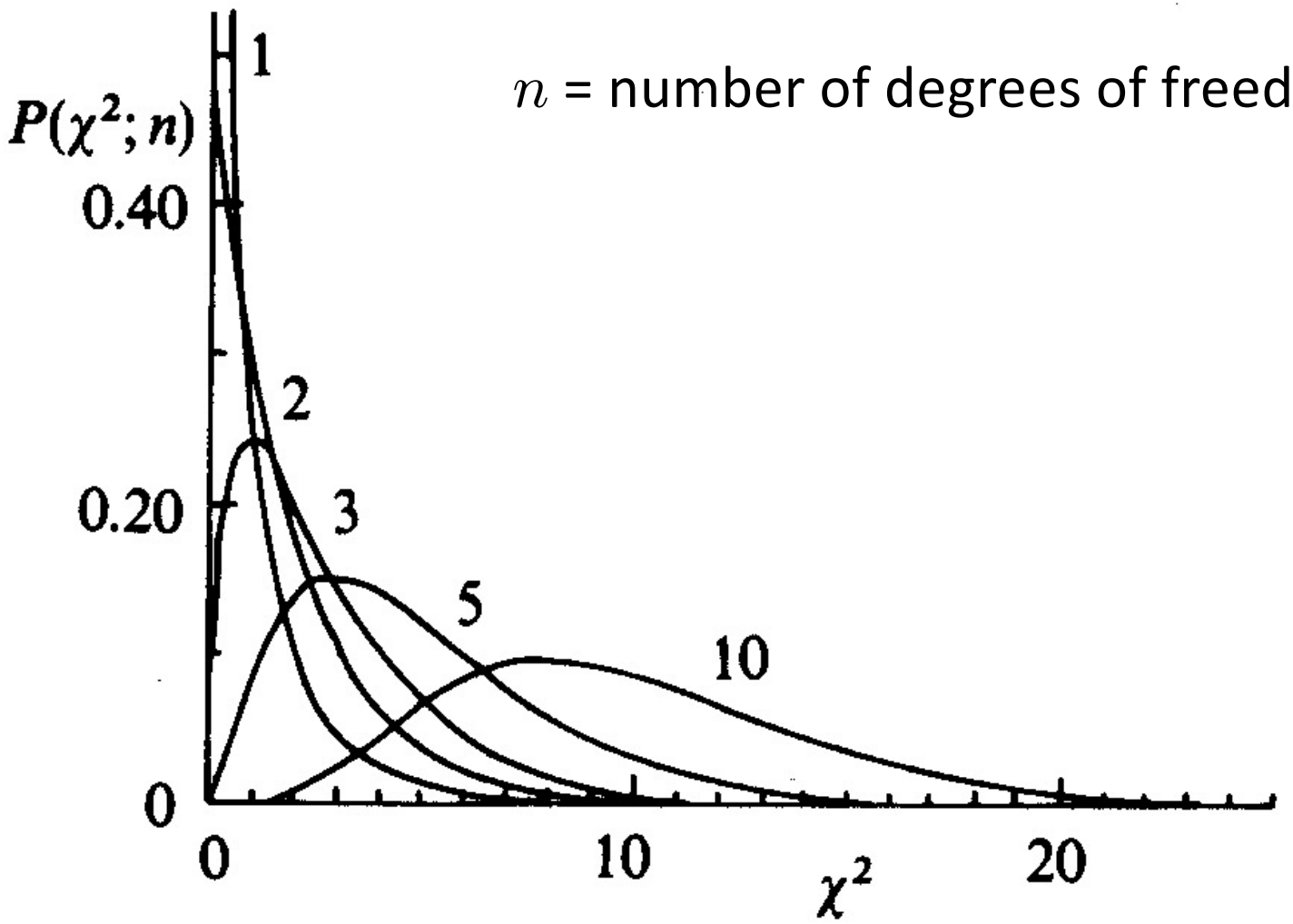
ML does not require rebinning, since it does not assume Gaussian PDF

Parameter estimation

The χ^2 procedure has the advantage that it not only provides the best-fitting parameters, but it also provides the **goodness of the fit**.

The reason is that the quantity χ^2 follows a chi-square distribution with $n = N - M$ degrees of freedom.

Parameter estimation



n = number of degrees of freedom

Parameter estimation

The χ^2 procedure has the advantage that it not only provides the best-fitting parameters, but it also provides the **goodness of the fit**.

The reason is that the quantity χ^2 follows a chi-square distribution with $N - M$ degrees of freedom.

The expected value of the chi-square distribution is $N - M$, and the variance is 2 ($N - M$).

Parameter estimation

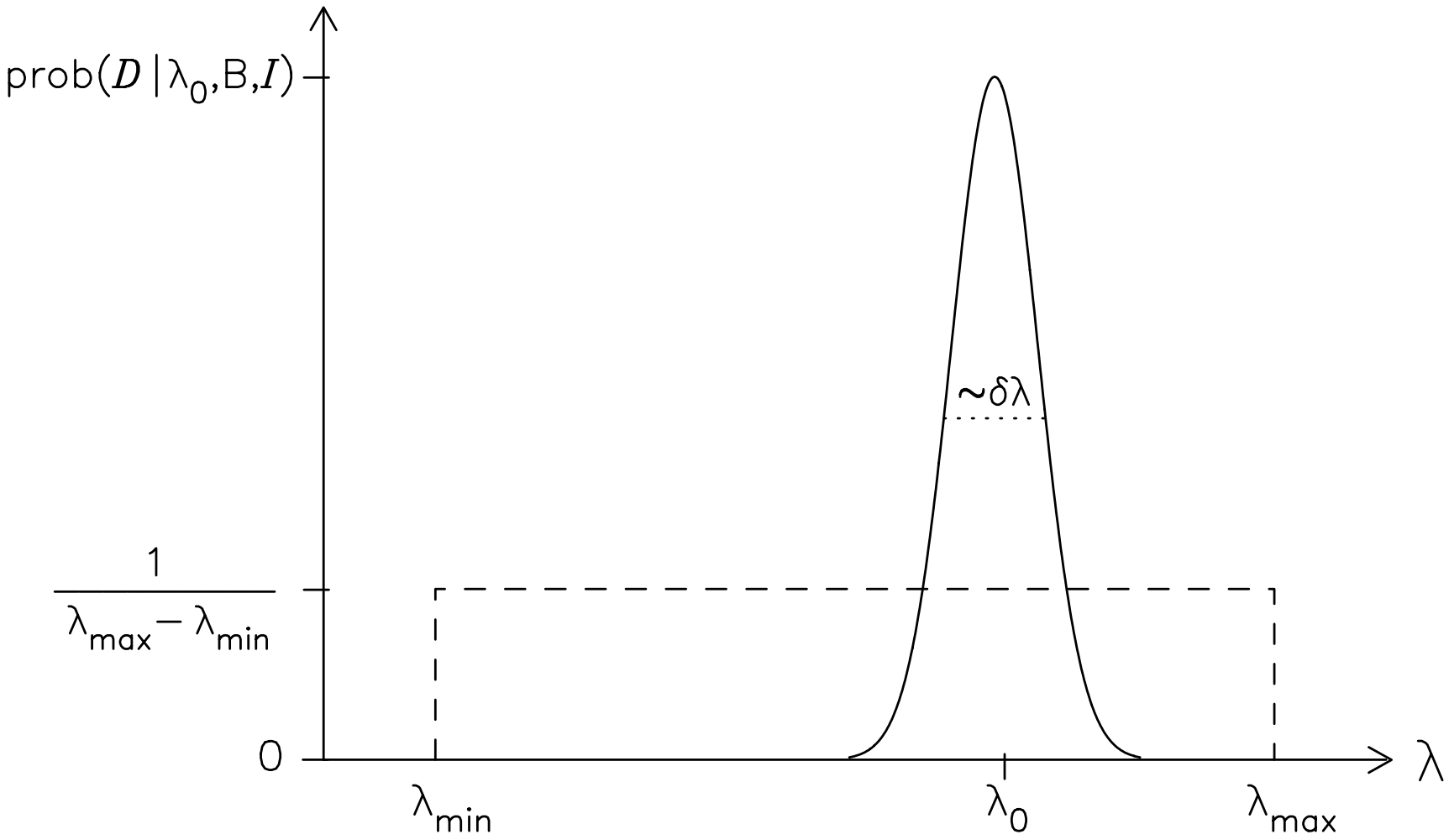
The likelihood gives the probability of getting the data given a model. But this is not what we want. We want the probability of a model given the data we have. Recalling Bayes theorem, the **posterior** is:

$$P[y(E_i;\boldsymbol{a})|\{y_i\}] \propto \mathcal{L}[\{y_i\} \mid y(E_i;\boldsymbol{a})] \times P[y(E_i;\boldsymbol{a})]$$

which is what we in reality have to maximize.

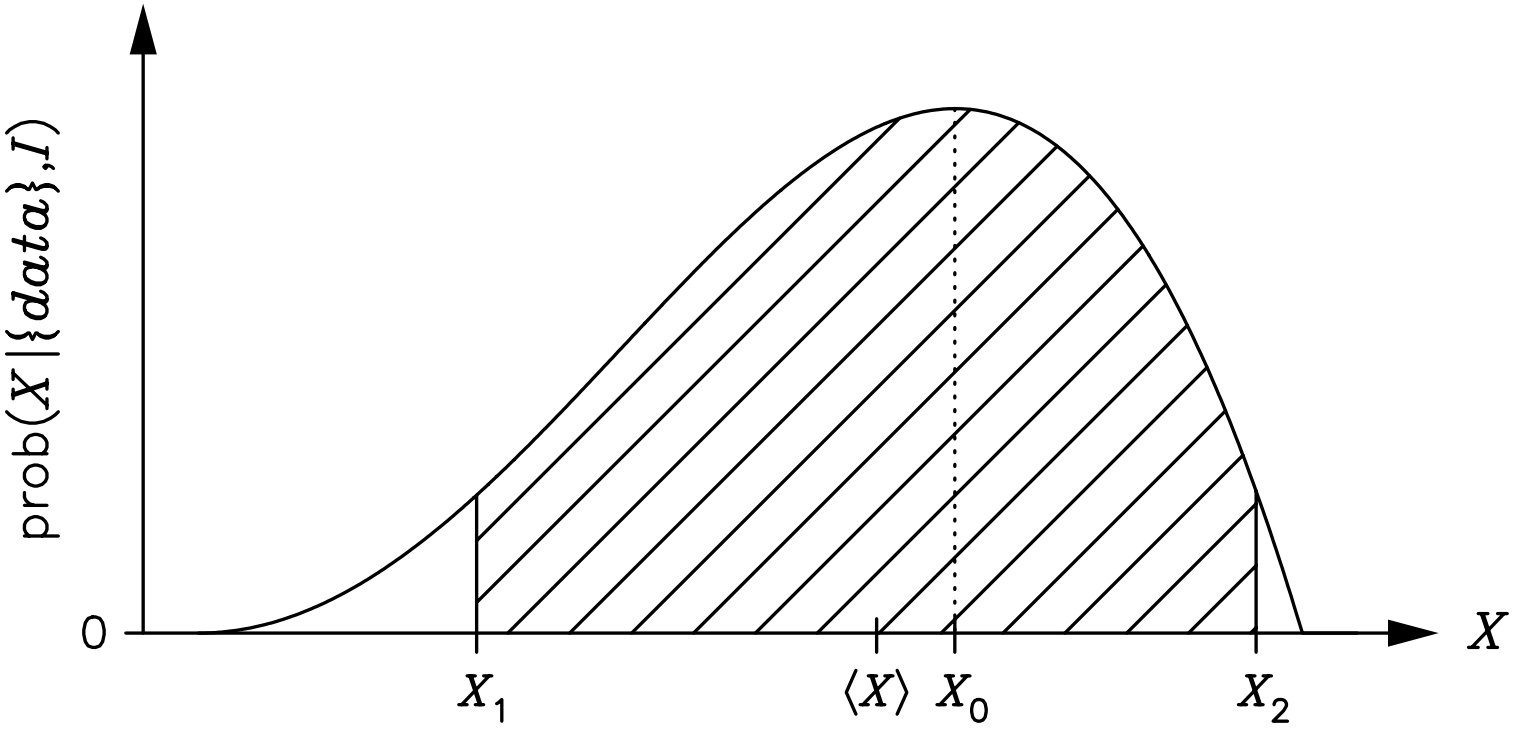
If we do not have any a priori information about the model (the parameters of the model), we can choose a uniform prior over a large enough range such that the relevant part of the likelihood is well within the range of the prior, and hence the **posterior** is simply proportional to the **likelihood**.

Parameter estimation



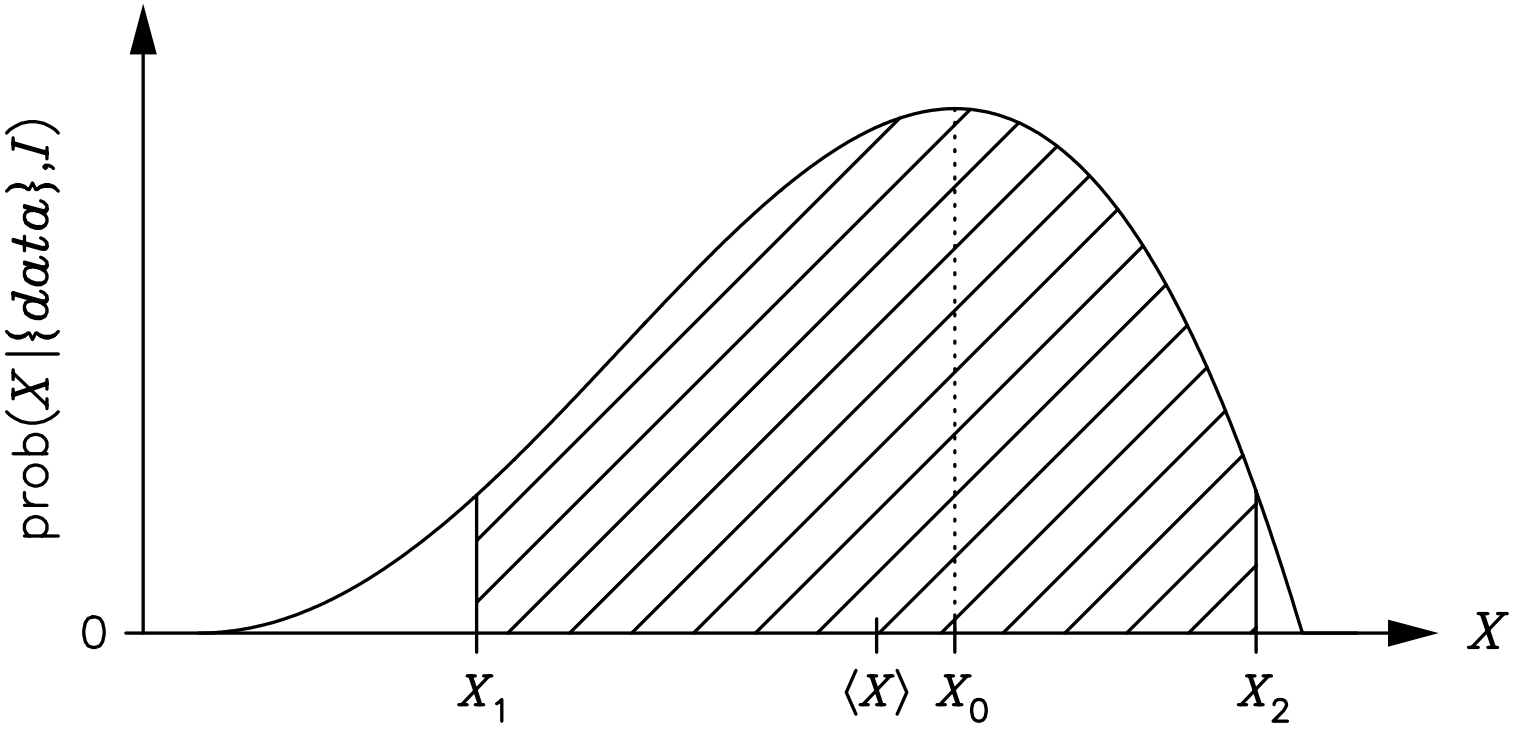
Confidence range

The posterior probability is all we need. From the posterior we can in principle find the best-fitting value of the parameters, and also the **confidence range**. The confidence range is the smallest interval of the posterior around the maximum that contains a given fraction (e.g., 68% or 90%) of the posterior.



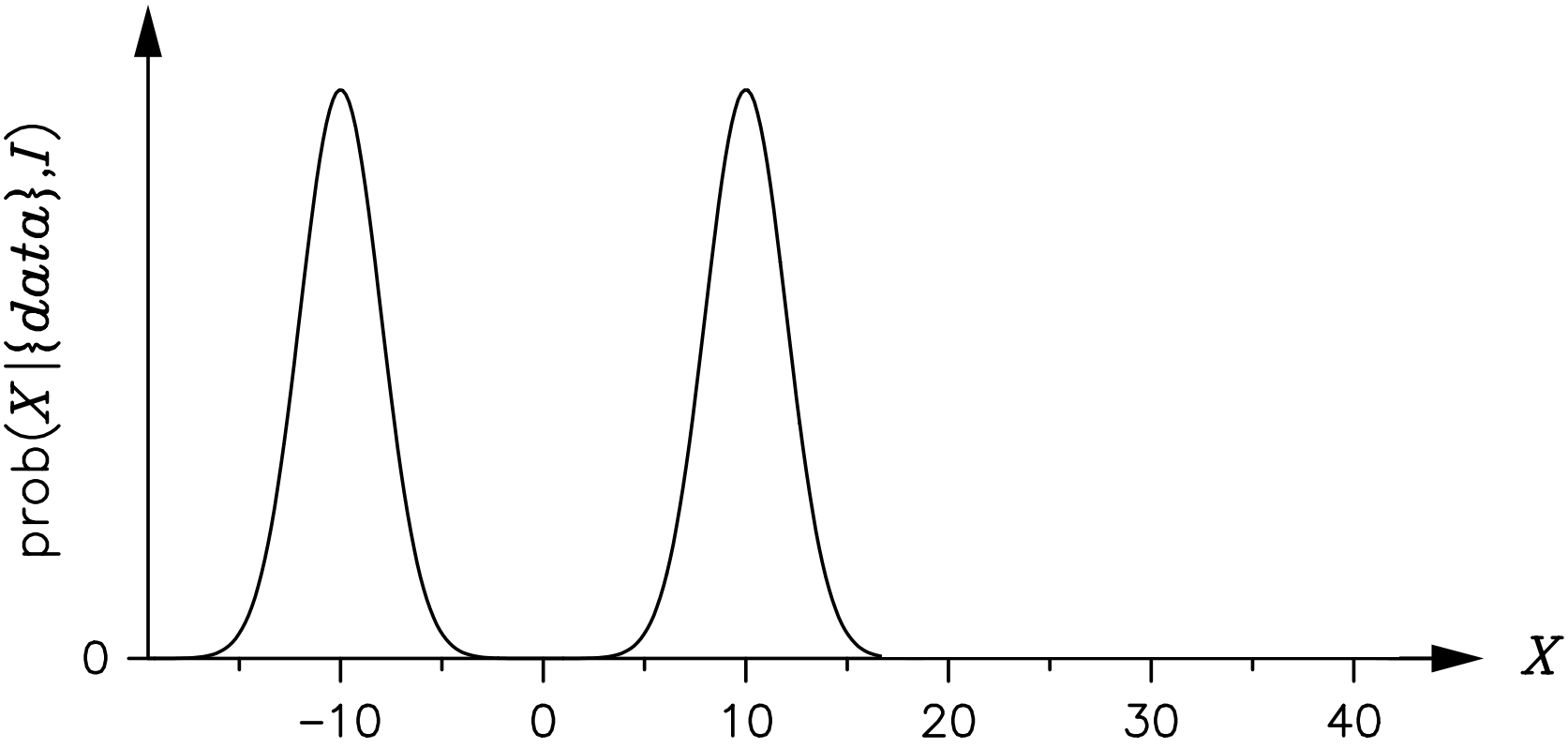
Confidence range

It is more common, however, to give the best-fitting value and the error (to a certain confidence level) of that best-fitting value. If the posterior PDF is asymmetric, it is customary to give separately the positive and negative errors.



Confidence range

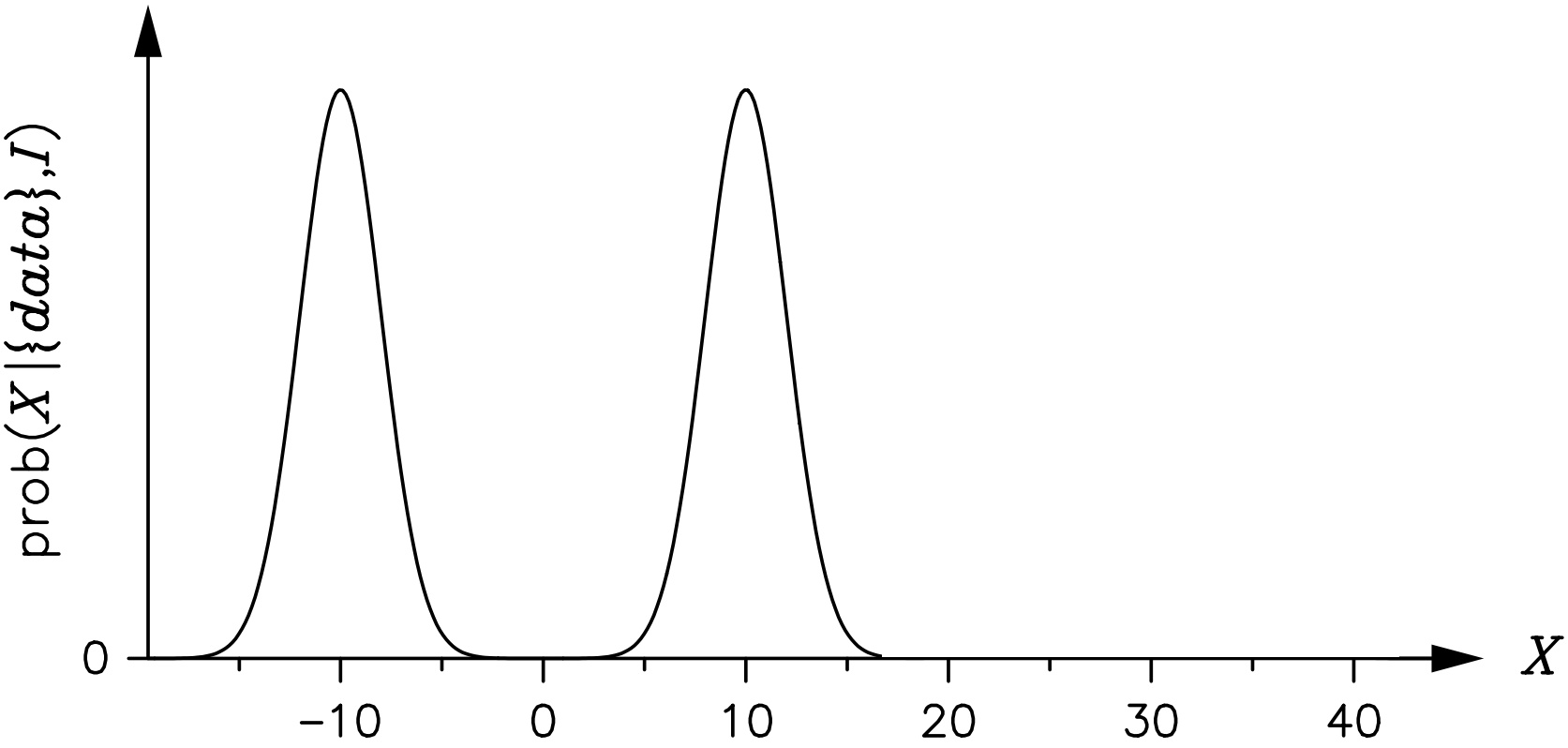
The posterior probability may be multi-peaked, with several of the peaks being more or less equally high (probable). This brings up the problem of how to quote the best-fitting value and the error (notice that the posterior PDF provides the right information!).



Confidence range

In this case, should one report $X = 0 \pm 10$? Notice that $X = 0$ has zero probability according to the posterior PDF!

Either give the full PDF, or report "$X = -10 \pm 1$ or $X = 10 \pm 1$"



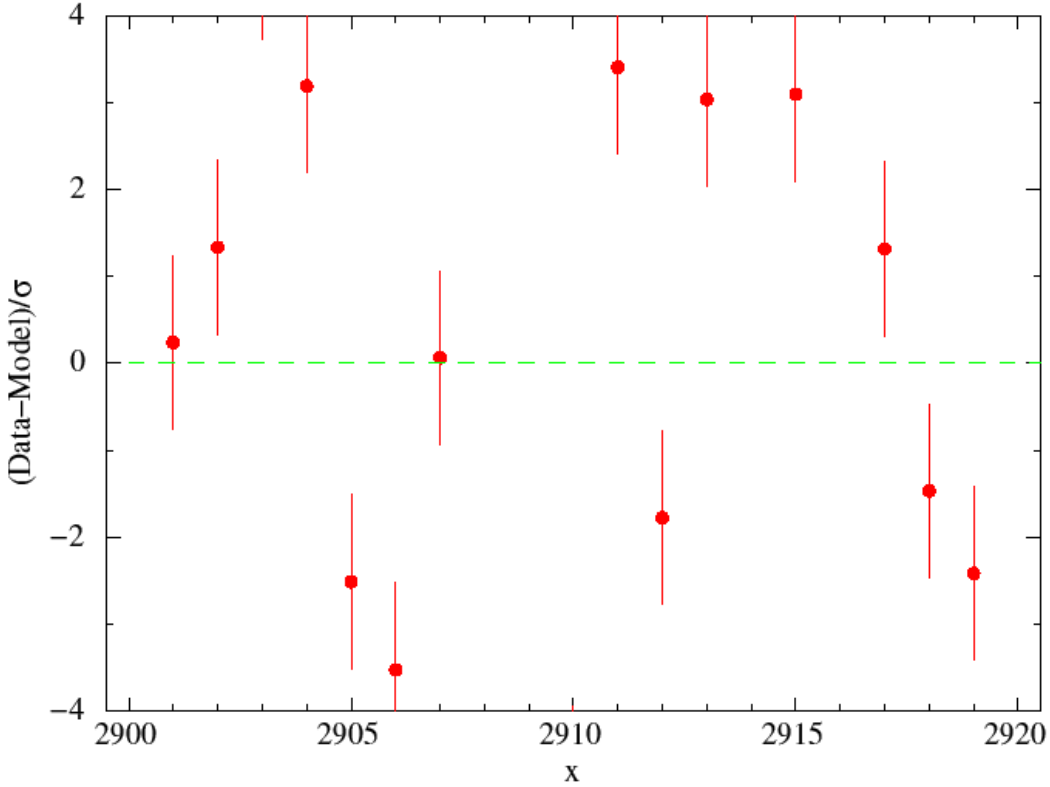
χ^2 – fit: Watch-out notes

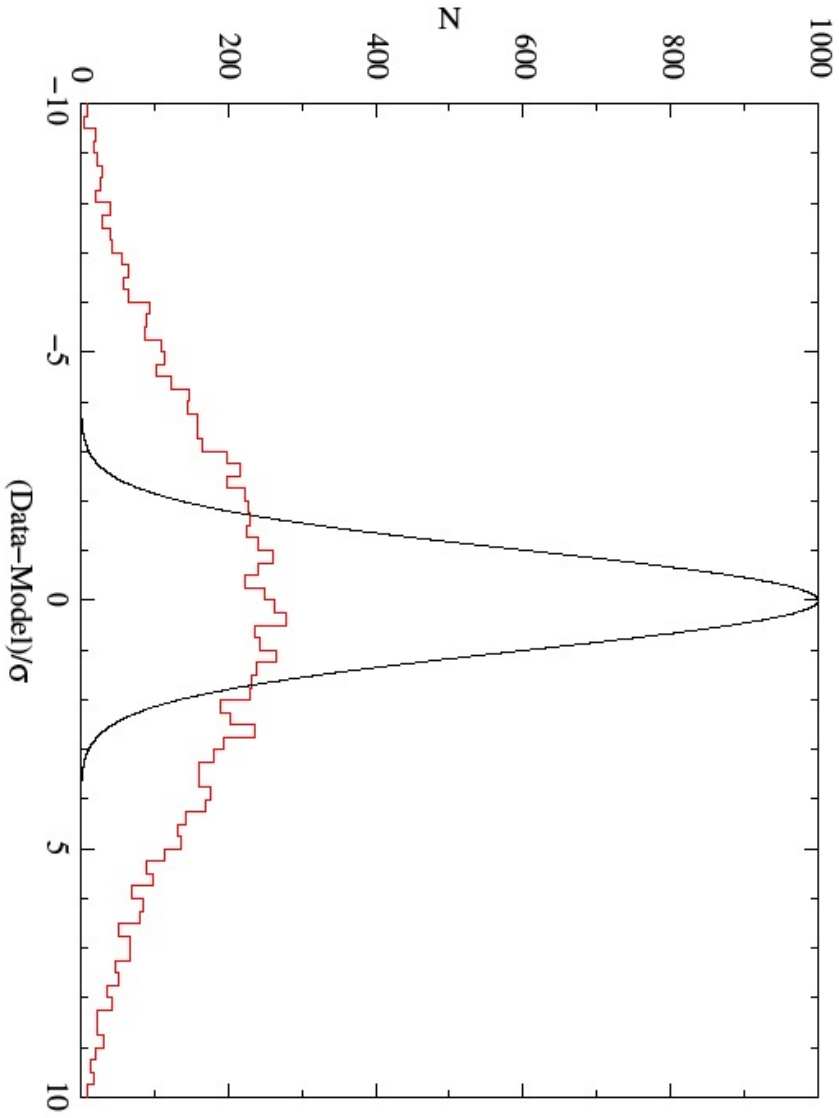
If you decide to minimize the χ^2, consider the following:

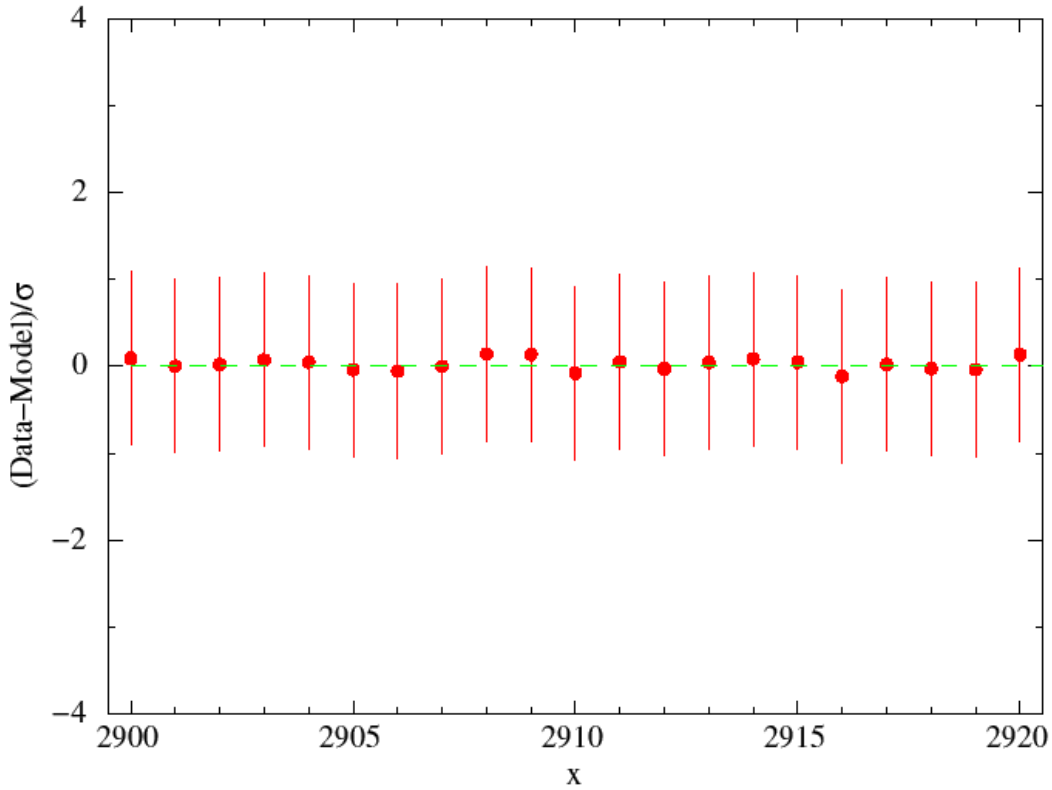
You have to find a model and parameters that make χ^2 approximately equal to the number of degrees of freedom, n (reduced $\chi^2 \approx 1$, where reduced χ^2 is χ^2/n).

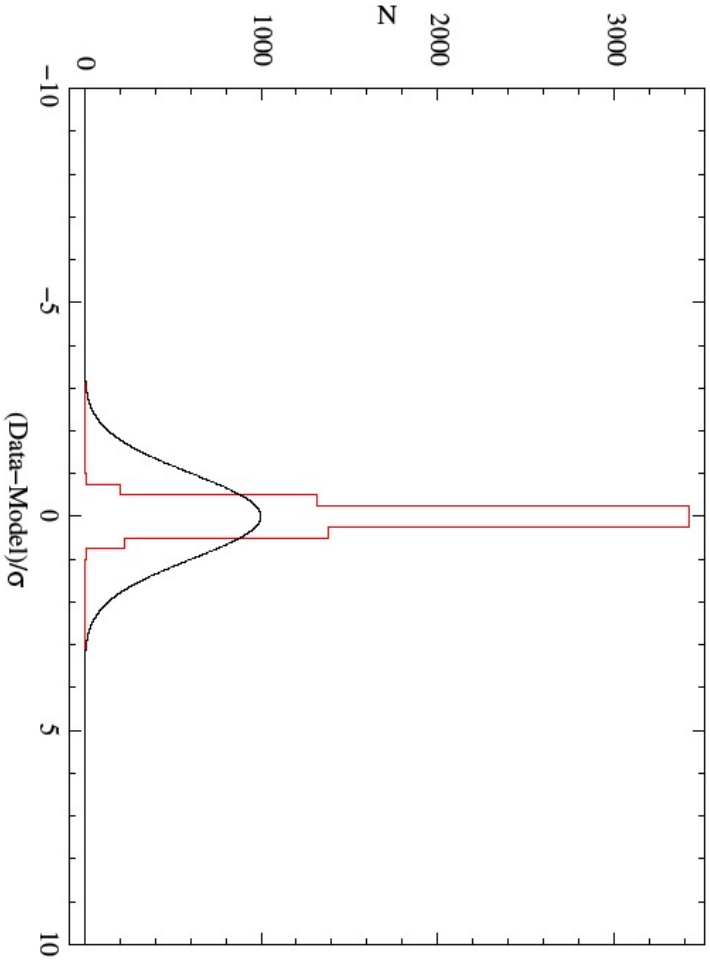
Against your intuition, a fit that yields a χ^2 close to 0 is worse than a fit with $\chi^2 \approx 1$.

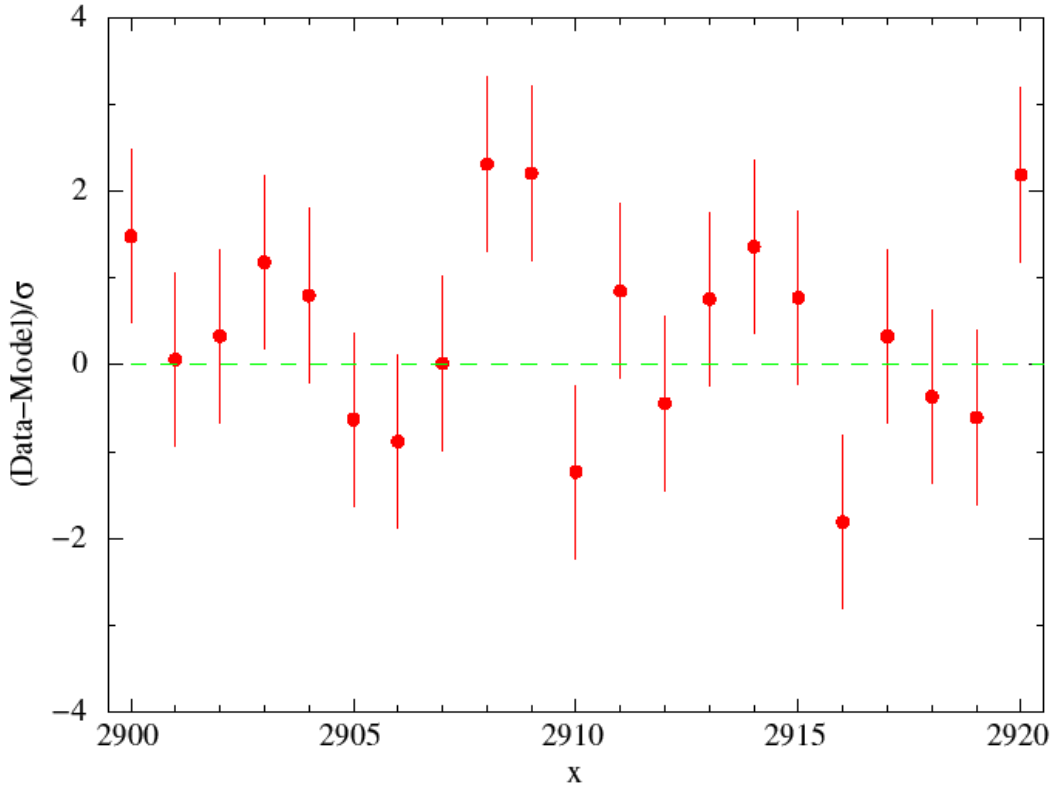
Do not try and keep adding parameters to reach $\chi^2 = 0$!!!!

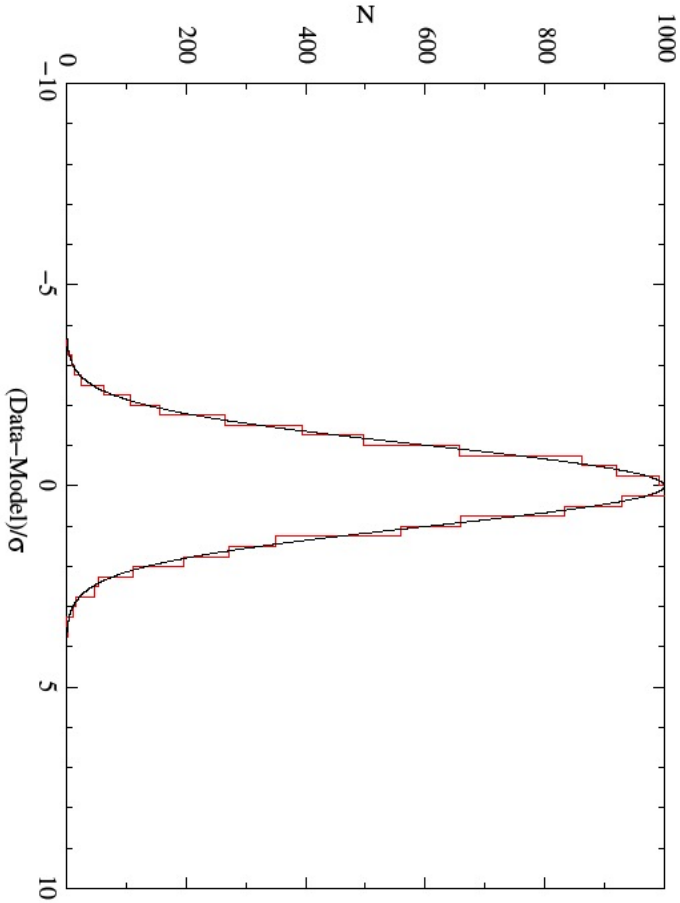

4
2
0
−2
−4
2900
2905
2910
2915
2920
x
(Data−Model)/σ


N
0
200
400
600
800
1000
−10
−5
0
5
10
(Data−Model)/σ


4
2
0
−2
−4
(Data−Model)/σ
2900
2905
2910
2915
2920
x


0
1000
N
2000
3000
−10
−5
0
5
10
(Data−Model)/σ


4
2
0
−2
−4
2900
2905
2910
2915
2920
x
(Data−Model)/σ


N
0
200
400
600
800
1000
−10
−5
0
5
10
(Data−Model)/σ

χ^2 – fit: Watch-out notes

Given the data, y_i, and the model at the same energy channels as the data, $y\,(E_i;\boldsymbol{a})$, we want to minimize:

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - y(E_i;\boldsymbol{a})]^2}{\sigma_i^2}$$

where $\sigma_i = y\,(E_i;\boldsymbol{a})^{1/2}$ (**expected error**).

However, since we neither know the model (that is what we are after) nor yet fitted the data, we do not know what is the expected value, $y\,(E_i;\boldsymbol{a})$, and hence we cannot calculate the expected error.

One normally takes the observed error, $\sigma_{i,\text{observed}} = y_i^{1/2}$, as a proxy to the expected error.

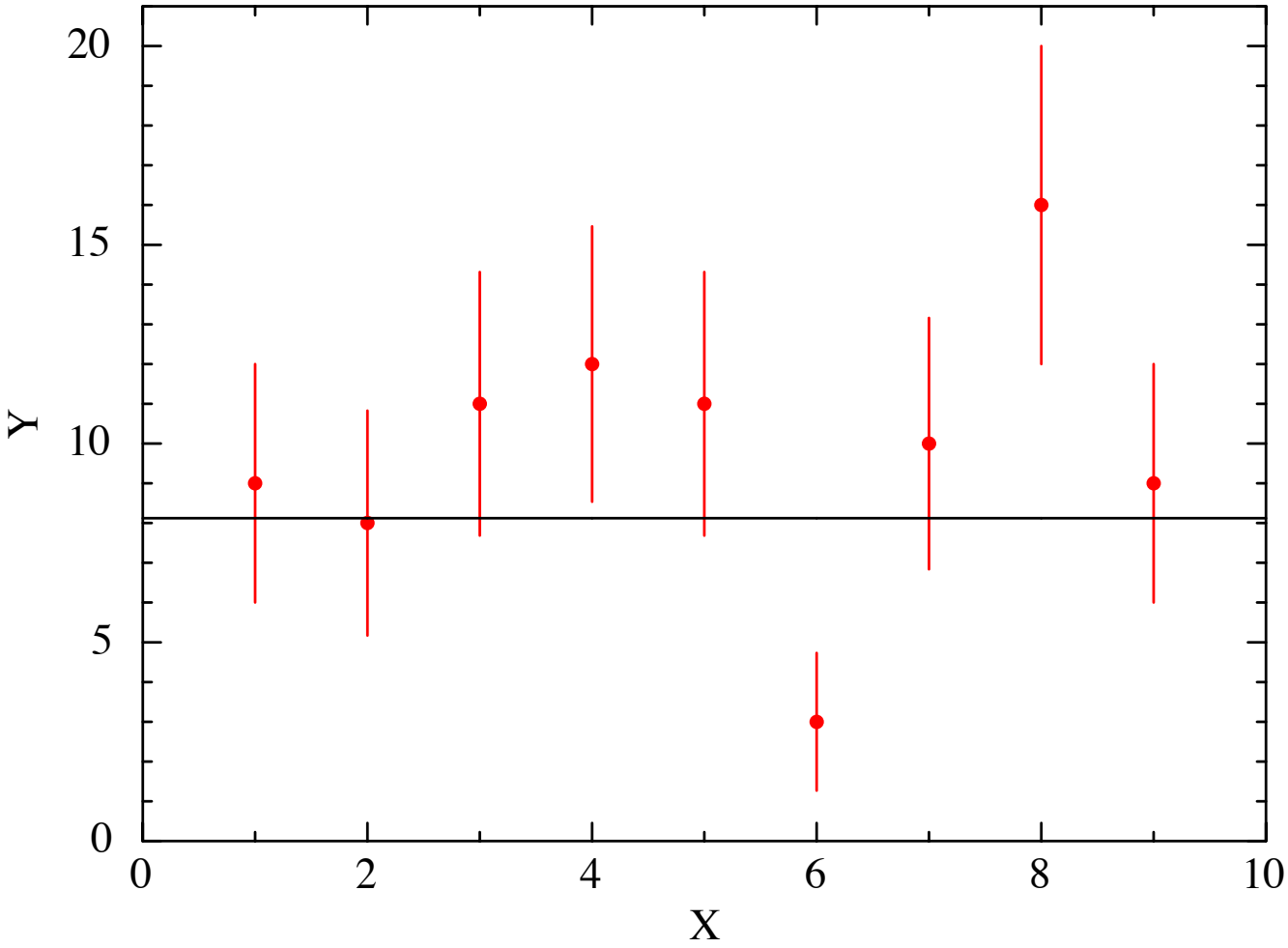
χ^2 – fit: Watch-out notes

Notice that this error is “biased” (it is not a proper representation of the true error).

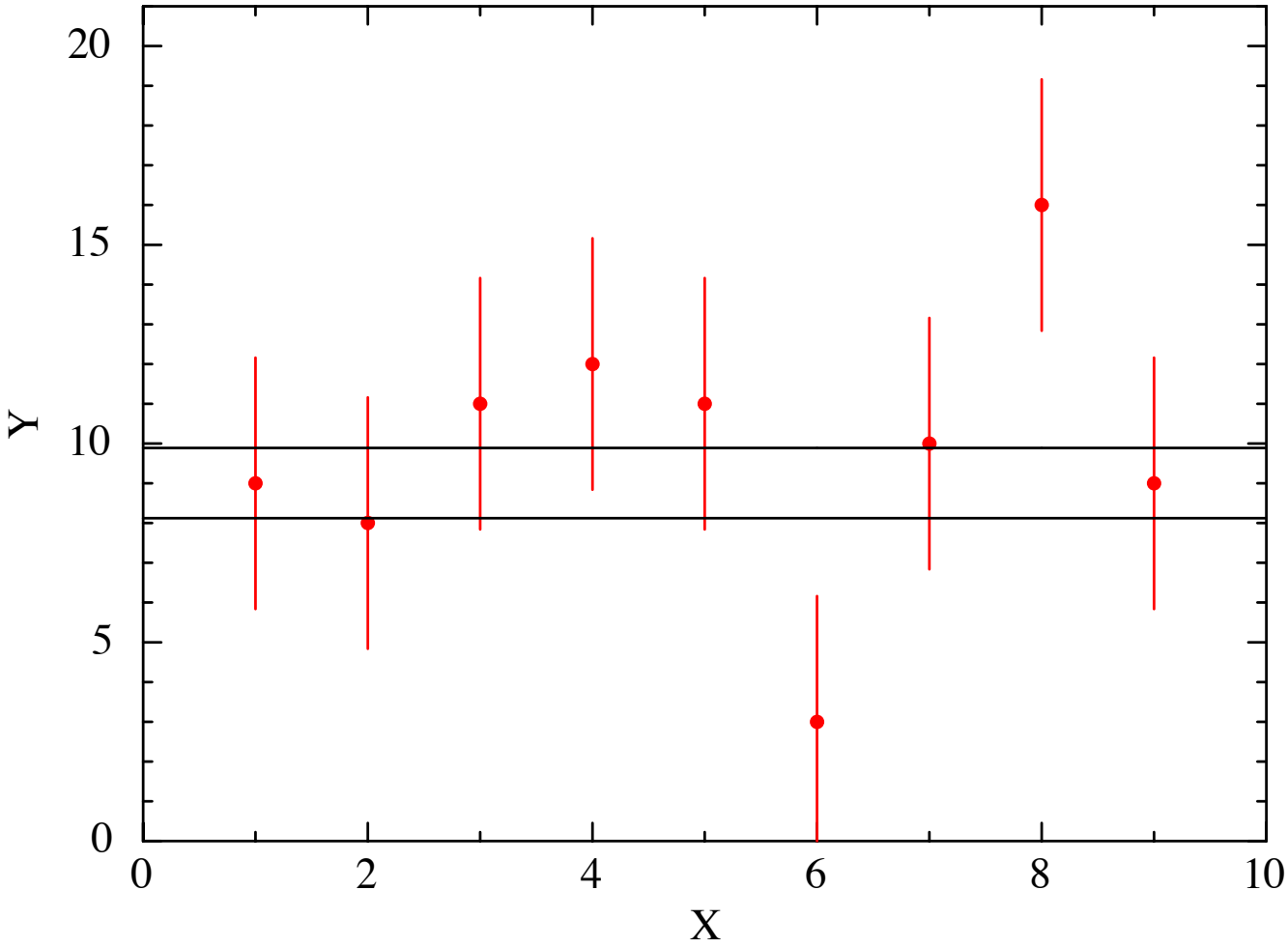
For instance, if one channel happens to have fewer photons than expected, the error will also be smaller than expected, and that channel will have a strong influence in the value of χ^2. In the extreme, if a channel has 0 counts, χ^2 goes to infinity!

Xspec offers a few solutions (see command “weight”).

χ^2 – fit: Watch-out notes



χ^2 – fit: Watch-out notes



Hypothesis testing

Once you have fitted your model, you want to know whether there is another model that would fit the data better.

For instance, what if I add a line to my model? Does the fit improve? Is the line significant? Should I add another line to improve the fit? When should I stop?

Or I fitted a power law to my spectrum. Do I get a better fit if I include a high-energy cut off?

This is a very common problem in science, and it is one that has been explored a lot. One can approach this from the Bayesian point of view, and compare the posteriors of the two models, but a quantitative assessment of the improvement is still missing, and is a topic of continuous studies.

Hypothesis testing

Under some circumstances, there is a ”frequentist” approach, the so-called **F-test** (developed by Fisher) that helps answer some of these questions.

Adding new parameters to the model (hopefully!) improves the fit at the expense of reducing the number of degrees of freedom (remember that number of degrees of freedom is the number of data points minus the number of parameters).

The idea is to compare the χ^2 and the number of degrees of freedom of the two fits, e.g. one with and the other without the line.
A combination of these 4 numbers follows a specific distribution, the **F distribution** for n_1 and n_2 degrees of freedom, where n_1 and n_2 are, respectively, the number of degrees of freedom of each fit.

Hypothesis testing

Xspec has a command called “ftest” that gives you the probability that the improvement in the fit happened only by chance.

If the probability is low, one can conclude that it is unlikely (but never certain!) that the improvement is not significant or, in other words, it is quite likely that the addition of the extra parameters improves the fit significantly.

Bear in mind that two conditions must be met in order to be able to apply the F-test (see Protassov et al. 2002 ApJ 571, 545):

(1) The models should be nested.

(2) The new model should not be equal to the old model at an extreme value of one of the parameters.

Hypothesis testing

(1) The new model becomes the old for some value of one of the parameters. For example:

$y_1(x) = a + b\,x$

$y_2(x) = a + b\,x + c\,x^2$

are nested because $y_2(x)$ becomes $y_1(x)$ for $c = 0$.

$y_1(x)$ = blackbody

$y_2(x)$ = powerlaw

are not; you cannot convert a power law into a blackbody for any value of the power-law normalization or power-law index.

Hypothesis testing

(2) The new model must not become the old one at an extreme value of one parameter. For example:

$y_1(x)$ = a + b x

$y_2(x)$ = a + $b\,x$ + $c\,x^2$

is okay if c can be either positive or negative because c=0 is not an extreme value of c, but:

$y_1(x)$ = powerlaw

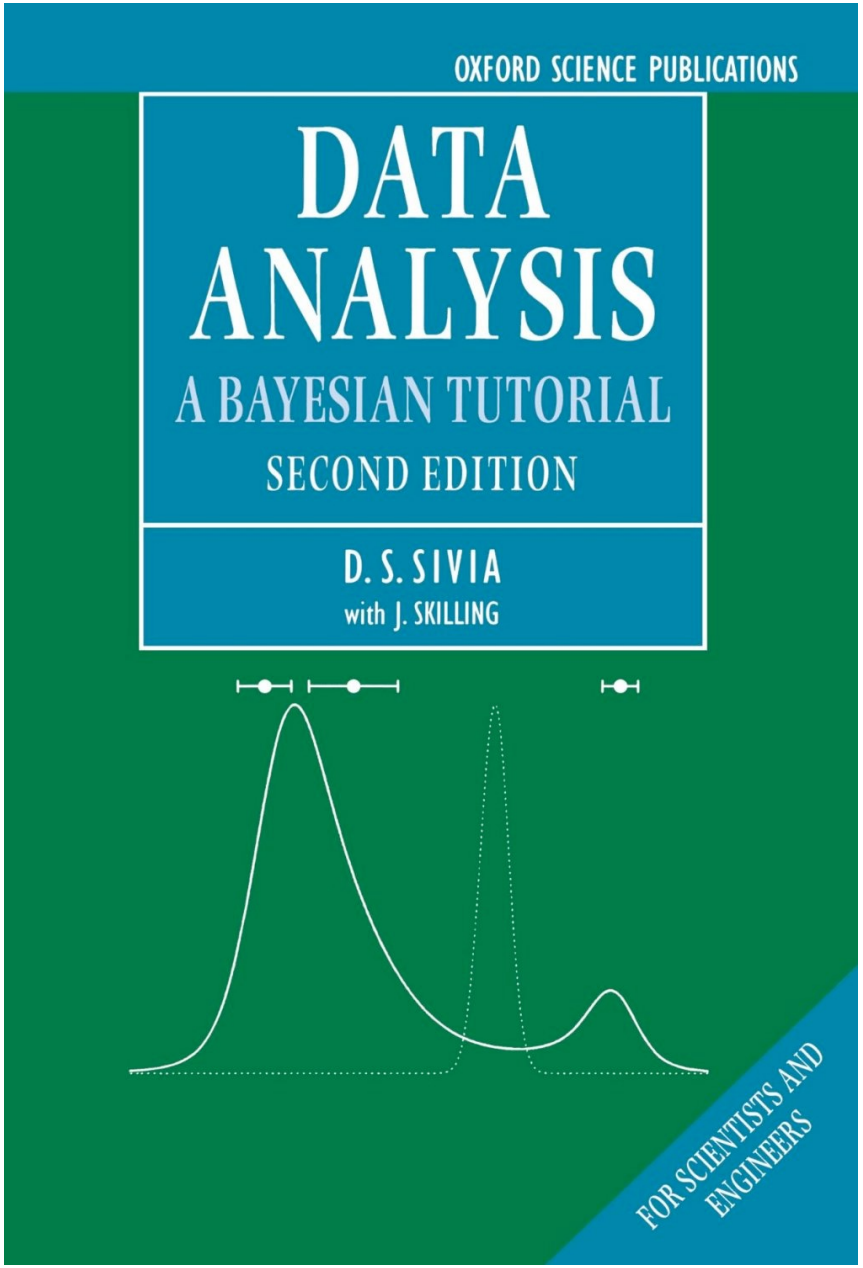
$y_2(x)$ = powerlaw + $A \times$ gaussian

is not okay if this is an emission line ($A \geq 0$), since $y_2(x)$ becomes $y_1(x)$ for the extreme value A = 0.

Hypothesis testing

If your case does not satisfy (1) and (2), you cannot be sure that the F-test gives you the right probability (it may, but it may not).

In that case you should use Montecarlo simulations.



Now, fit and enjoy!